Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 12:40:04
REFERENCE NO: 236

Author Names & Affiliations

- Gary Holton - University of Hawaii at M?noa

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Linguistics

Title of Submission

Addressing the collection management bottleneck to facilitate a more data-driven science of linguistics

Abstract (maximum ~200 words).

The challenge of documenting the world's linguistic diversity has been answered by programmatic responses in the field of linguistics, leading the emergence of dedicated academic training, funding schemes, archiving infrastructure, and best practices for language documentation. This ramping up of capacity took place during the past two decades during a time of both social and technological upheaval. Social changes including rapid urbanization have led to a dramatic increase in rates of language death, forcing linguists to work ever faster to document linguistic diversity. At the same time, shifts in technology – particularly the transition from analog to digital recording techniques -- have completely rewired the toolset for the field linguist.

Today a significant bottleneck remains in moving linguistic data from the point of collection to a data repository where it can be accessed by a wider user community. Previous efforts have focused on either the extreme ends of the data management continuum. Tools for the field worker are now well-advanced, and at the top level there now exists a well-developed network of linguistic data repositories. Moving data from the field to the archive remains a significant challenge, owing to a lack of standards and tools.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The recent surge in efforts to document the world's more than 6000 languages has resulted in vast trove of natural language data, offering the potential for a more data-driven science of linguistics. However, realizing this potential has been hampered by the lack of tools for managing individual collections and transferring those collections to data repositories. Good tools now do exist for data collection by field linguists; and data repositories have been developed through a network of language archives (DELAMAN). Yet the pathway from data collection to data archiving remains fraught, to the point the many linguists continue to hoard data in private collections, even though they

may support the principles of open data and data sharing.

Thanks to new digital tools for linguistic fieldwork, linguists now routinely create tens of thousands of data files, including audio recordings, video recordings, photographs, annotation files, lexical database files, stimuli responses, etc. Different tools are used to work with different parts of the data, so that maintaining the relationships between the various parts of the data set can be challenging. This is further complicated by the fact that much of the work of analyzing and annotating these data is now done collaboratively by teams of linguists working together with language community members. Version control can be particularly difficult when collaborators are distributed across space and time. Linguistic annotation and analysis is an iterative process, requiring months or even years of work. Thus, the risk of data loss due to mismanagement of digital files is extremely high.

Some limited tools for linguistic data management do exist, but they have not been widely adopted. Arbil (Withers 2012) is a tool that was developed specifically for the DOBES project to support IMDI metadata, but has limited uptake outside the (now defunct) DOBES community, in part due to its non-intuitive user interface (Define 2014). Similarly,SayMore (Hatton 2013) was created for SIL International and has not seen much use outside that community. Most field linguists continue to make use of ad-hoc, idiosyncratic approaches to data management. Crucially, there is NO STANDARD OF PRACTICE in the field. Some linguists have been experimenting with the use of Git and other version control systems, but linguistic data provide particular challenges for conflict resolution between versions, due to the distributed nature of many linguistic data file formats (e.g., ELAN transcription files). Hence, a more robust solution is needed.

Ideally, linguists would archive data in repositories directly from the field (Robinson 2006). But the vast amounts of data now being generated by linguistic research teams make this increasingly difficult. The delay in processing prior to archiving can extend several years or more. Linguistic archives and funding agencies report low levels of compliance with requirements to deposit data. And typically only a fraction of collected material is deposited. This may be due to some intransigence on the part of the linguists, but the disorganized state of individual data collections is also a major barrier. Moreover, it is this barrier which can be overcome with appropriate technology.

We stand at the cusp of a new kind of linguistics which allows us to draw insights from a vast set of data on human languages. In the past linguists have faced a methodological tradeoff between depth of knowledge and breadth. On the one hand, they could draw insights from a very detailed first-hand knowledge of one or a handful of languages. On the other, they could make broad typological conclusions based on a superficial sampling of secondary data. Crucially, typological insights have been derivative, relying not on access to primary data but rather on the (often unverifiable) generalizations of others. A data-driven science of linguistics allows direct analysis of primary data across vast numbers of languages.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

This is not a technically difficult problem to solve. We have good models from other fields which routinely manage vast amounts of data. It is very likely that the necessary CI for this already exists outside of linguistics. So there is unlikely to be a lot of “innovative” science involved in adapting that CI for linguistics. But this is part of the challenge. Even though the CI itself may not be particularly innovative, a collection management framework for linguistics will facilitate an entirely novel and innovative approach to understanding human language.

Consent Statement

- “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”